Using APIs for data collection on social media

Lomborg, Stine; Bechmann, Anja

Published in:
The Information Society

DOI:
10.1080/01972243.2014.915276

Publication date:
2014

Document version
Early version, also known as pre-print

Citation for published version (APA):
Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*(4), 256-265. https://doi.org/10.1080/01972243.2014.915276

Lomborg, S: & Bechmann, A. (2014) Using APIs for data collection on social media. *The Information Society, vol. 30*(4) [Post-print version].

Using APIs for data collection on social media

Abstract

This article discusses how social media research may benefit from social media companies making data available to researchers through their Application Programming Interfaces (APIs). An API is a back-end interface through which third party developers may connect new add-ons to an existing service. The API is also an interface for researchers to collect data off a given social media service for empirical analysis. Presenting a critical methodological discussion of the opportunities and challenges associated with quantitative and qualitative social media research based on APIs, this article highlights a number of general methodological issues to be dealt with when collecting and assessing data through APIs. The article further discusses the legal and ethical implications of empirical research using APIs for data collection.

Keywords: APIs, social media, data collection, data quality, methodology, research ethics.

Introduction

Social media increasingly pervade everyday life in the developed parts of the world, enabling communication among users and collecting massive amounts of data for social media companies to refine and monetize their products. As part of their business model, social media companies often make their APIs (Application Programming Interfaces) available to third parties. An API is basically an interface of a computer program that allows the software to 'speak' with other software. This enables the development and enhancement of the core social

media services, for example, by allowing for third-party companies to develop their own software clients for using Twitter or integrating Facebook with other social media services.

Some social media companies make their data banks on users and usage patterns available through their APIs. Hence, the API is also an interface for researchers to collect data from a given social media service. Through small software scripts, researchers can access the API to retrieve, store, and manipulate digital traces left by the users of a service for further empirical analysis.

APIs present significant opportunities for internet research of both quantitative and qualitative nature, and we are only beginning to see the results of such research endeavors. Part of the allure of API-based research is that the collection, organization, cleaning, preservation and analysis of data can be automated, thus making APIs highly efficient research tools. Along with web crawlers, network analysis software and similar digital research tools and techniques, the use of APIs to collect data is linked to what is labeled 'the computational turn' in social science (Berry 2011; also boyd and Crawford 2012). The ability to collect the digital footprints of a person is compelling from an internet research point of view as it potentially means that we can get access to usage patterns (communication, connectivity etc.) in greater detail and more systematic fashion than other methods such as interviews and surveys allow for. Moreover, the methods for collecting the digital footprints are instantaneous and non-intrusive. At the same time, as it will be argued in the following, it is not clear to which extent the APIs of social media are actually 'open' for researchers in the sense of offering valid and reliable access points for collecting empirical

data. In this article, we examine the use of API's as one of the key trajectories of computational, empirical scholarship on the uses of social media.

In contrast to web crawling, which can traverse only the public web, the use of APIs provides access to non-public internet environments, such as those requiring authentication through login and password, because the data collection runs directly through the back-end of the social media service to which the data belong. In comparison to the use of raw log files that require formalized collaboration or affiliation with the social media company who owns the data, APIs are publicly available. Instead of having to involve the company to get them to deliver data, through the API researchers can themselves collect relevant data from the data pond made available by the company. However, whereas raw log files can reliably deliver all data for all users to the researcher (e.g., including time spent on the service, privacy settings, friending history, and the click-behavioural patterns of the users), API-based research suffers from a lack of transparency regarding the data output and quality, which may significantly weaken the research. To be sure, while social media APIs make data publicly available, they are not open in the sense of giving full and unlimited access to the entire database of companies such as Twitter and Facebook.

In this paper we aim to critically review and discuss the opportunities and challenges of using APIs for researchers wanting to study the behavioral patterns of internet users in and across social media services, whether these are publicly accessible, or by default require user authentication and allow for the users to further install privacy measures to protect their profiles from public scrutiny. In other words, *what are the methodological challenges of getting access to and*

analyzing data on social media through APIs, and how can these challenges be dealt with?

To address these questions, first, we provide a brief review of social media research that uses APIs as tools for data gathering, to indicate the span of opportunities for working with this kind of data. Whereas most existing research is quantitative, we suggest that there is ample potential for qualitative research as well. We then map and critically discuss the territory of methodological challenges for using APIs in social media research and discuss the broader context of legal and ethical issues involved when studying social media through APIs in a European context.

To exemplify the methodological challenges and ethical issues we use Facebook and Twitter as key instances of social media that make data available for researchers through their APIs. Facebook and Twitter are the most widely diffused social media in an international perspective (Serrano 2011).[1] The two services differ in an important respect with regards to data retrieval from the API. Twitter is public by default and profiles are accessible also to non-members of the service. Facebook is for a large part a fenced-off service. That is, while some profiles are publicly accessible, many users deploy various privacy settings offered by Facebook to restrict access to their profiles. Whereas many challenges of using APIs for research are similar across services, there are differences between social media services in the way their APIs are structured and what they give access to. The unique challenges of Facebook and Twitter are to some extent reflected in this article's main points. However, we try to keep the analysis on an abstract level that comprises the general use of APIs for empirical social media research.

Overview: Existing studies using APIs as a research tool

A research strategy that is currently in the vogue is the use of APIs to extract large amounts of behavioral data on social media use. In quantitative research, this involves, for instance, harvesting and analyzing so-called 'big data' (e.g. Bollier 2010; boyd and Crawford 2012; Rogers 2012). Big data is typically associated with massive-scale logs of user behavior in digital systems, and these data are extracted and mined in terms of pattern recognition, detection of deviant patterns that need further attention, and to develop predictive models (Vicente, Assent and Jensen 2011). In qualitative research, APIs are typically used to harvest textual archives of communication patterns on social media for close-up analysis. In this section, we review existing social media research that relies on APIs for data collection. To date, most API-based research is carried out on Twitter. Historically, Twitter has been quite open in terms of access to its database, although the types of API-based access have become much more restricted over time (González-Bailón et al. 2012).

A popular type of analysis that can be performed on relational data from log files extracted through APIs is network analysis, and its associated compelling visualizations of relationships among nodes in structural patterns. For example, Huberman, Romero and Wu (2009), collected data from more than 300,000 Twitter users to study relationship patterns on Twitter, including the distribution of profiles according to the number of followers, the follower/following ratios, and the directionality of communication (i.e. whether or not tweets were part of a conversation). Wu and colleagues (2011) collected a

five billion sample of tweets in order to examine the use of links on Twitter. Moreover, network analysis based on API data has been used to visualize thematic threads in the network and analyze the social connections and/or the diffusion of tweets (e.g., Ausserhofer and Maireder 2013; Bruns and Burgess 2012). While network analysis can also assist qualitative research (e.g. Lomborg 2011; Markham 2012), it is particularly valuable as a tool to mine and visualize patterns from large sets of relational data (Hansen, Shneiderman and Smith 2011).

Another main trajectory of API research draws on techniques derived from content analysis to examine semantic patterns in social media communication. Java and colleagues (2007), and later, Naaman, Boase and Lai (2010) collected samples of more than 100,000 public tweets and performed a content-based categorization of Twitter messages to distinguish between different communicative purposes and estimate their relative prominence. Horan (2012) studied forms of produsage in the relationship between soft and hard news on Twitter using a random sample of more than three million tweets collected over a week through the API. Weller and Puschmann (2011) used the Twitter API to collect scientific conference hashtags and scientists' Twitter profiles in their study of citation patterns through the use of URL's in online science communication. Similarly, Lotan and colleagues (2011) queried the Twitter API for tweets containing specific hashtags related to the Tunisian and Egyptian uprisings in early 2011, as well as public profile information for the sampled tweets, which was used to classify the actors involved in disseminating and sourcing information about the uprisings on Twitter. Other studies examine the function of specific features of Twitter. Honeycutt and Herring (2009) studied

Twitter conversations with specific emphasis on the @reply sign as a marker of addressivity and interactional coherence between tweets.

In addition to such studies collecting data directly from APIs, APIs have also been used as a sampling tool for survey-based studies of social media. For instance, in their study of uploaders' expectancies and audience feedback, Courtois, Mechant and Marez (2011) used the Google data API to sample YouTube users who had uploaded content to the service.

Most social media research using APIs relies on publicly accessible data that is accumulated on social media services, typically using specialized software scripts to access, collect and process sampled data from the API. In contrast to API-based research, a few studies examine total populations of given social media services. For example, Kwak and colleagues (2010) crawled all Twitter profiles in June 2009 in a study of follower/followee patterns, activity levels, trending topics and retweets. Using raw server log files, rather than the API, Leskovec and Horwitz (2007) examined MSN on a planetary scale and analyzed user demographics and geographical locations, communication patterns and network homophily. In addition to this, social media companies themselves conduct raw logfile analysis to optimize data presentation and personal targeting, although this research is rarely reported in academic journals and therefore remains a black box for the academic community (Neuhaus and Webmoor 2012; exceptions include, e.g., Backstrom, Boldi, Rosa, Ugander and Vigna, 2011; Bakshy, Rosenn, Marlow and Adamic 2012; Bernstein, Bakshy, Burke and Karrer 2013; Kramer 2012; also Facebook's own Data Science Notes: https://www.facebook.com/data?sk=notes).

Most API-based research is quantitative. Despite an obvious potential, APIs as a tool for data gathering is quite rare in qualitative research. The use of APIs in these qualitative studies differs from quantitative studies, which collect data in structured databases and include a large number data points. In contrast, for qualitative research, the data collected through the API is typically structured and displayed as a textual or graphical archive (Lomborg 2012).

Lomborg (2011) studied the organization and norms guiding communicative practices on Twitter (e.g. conversational structures, thematic relevance and social negotiation and regulation of appropriateness). Through the API, she gathered textual archives of naturally occurring communication as well as meta-data (e.g., date and time stamps, message urls, user-ids of the users involved in communicating) from a purposive sample of Twitter profiles (N=6) for one month. The data was collected and organized in a database of communicative practices on Twitter using a customized script for TATIANA, an open source trace analysis tool for interaction analysis.[2] Bechmann (2014, forthcoming) studied how users connect online services together through Facebook and share data across these services. To study personal digital data traces the system Digital Footprints[3] was developed and used to access the Facebook API and collect data from a purposive sample (N=17) of high school students in Denmark and US. Bechmann gathered user participation data (e.g. profile data, walls, newsfeeds, and group data) as well as meta-data (e.g., date and time stamps, the url of the message, the user-id of the users involved in communicating) from when participants joined Facebook (and newsfeed for a period of 14 days).[4] In contrast to the Twitter API using the API of Facebook

means that researchers have to request permission to collect non-public data from the participants through a Facebook app.

Collecting, structuring and analyzing data from APIs for quantitative as well as qualitative research are associated with a number of challenges. For one thing, researchers need to have computational skills or collaborate with researchers who do in order to use APIs for collecting social media data. Furthermore, basic knowledge of different types of software, their strengths and weaknesses is required for selecting the relevant tools. While a substantial discussion of variations and biases in data collection software is beyond the scope of this article, it is important to note that the specific software used as an interface between the API and the researcher influences what kind of research questions that can be asked.

At the time of writing there are only a few free, open source tools for capturing data from APIs and 'cleaning' the data (e.g. OpenRefine (2012), Yahoo! Pipes (2007), iScience Maps (Reips and Garaizar 2011) and YourTwapperkeeper (2011)) and for processing, analyzing, and visualizing data (e.g., Gawk (2011), Node XL (Smith et al, 2009) and Gephi (2011)) (Bruns and Liang 2012). In addition to open source tools, many research groups at universities are developing their own tools and platforms (e.g., *Pattern* at the University of Antwerp (De Smedt and Daelemans 2012); *Digital Footprints* at Aarhus University (Bechmann and Vahlstrup 2013), and *Twitter Zombie* at Drexel University (Black et al. 2012), to mention only a few) customized to specific types of research questions. Furthermore, there is a growing industry of so-called social media analytics platforms (e.g. Radian6, Sysomos, DataSift) that offer paid tools for research. Such tools are an important part of working with

API data. These raise a number of additional issues for researchers, including an increasing need for funding and for collaboration with companies. Many of these issues are of a more practical nature, although they intersect with general questions of methodology, legal considerations and ethics, which are discussed in the subsequent sections.

Methodological issues: generalization, validity and reliability

In this section, we address a set of key methodological challenges of using APIs for research by considering the foundational issues of validity, reliability and generalization.

Sampling and constraints on generalization: Access to what and whom?

Twitter, Facebook and other social media companies' way of structuring their APIs set limitations on how researchers can construct their studies and what they can investigate through APIs as methodological tool (Bechmann and Vahlstrup 2013). In order to assess the value of data from APIs in empirical research, it is necessary to consider what types of user data it is possible to collect – and by extension, what kinds of users APIs can help us study.

In Bechmann and Lomborg (2012) we argue that social media research may benefit from a nuanced and critical reflection on the different actor roles that users may take on in their engagement with social media services – for instance, a small portion of users engage the Developer Toolkit on Facebook to create applications, software, and websites outside Facebook and connect them with Facebook, whereas other, larger groups of users post content and play games. Many simply connect but never participate actively by posting, and

others again barely log onto the site. Hence, social media users are no homogenous group, and the use of APIs to harvest behavioral data from social media has significant limitations in terms of accounting for the multifarious user roles.

Researchers can collect data from APIs about location, demographic information, newsfeed, uploaded material, the social graph, and so on. Evidently, the users who are most likely to generate most of these data in the systems are hardly representative of the entire population of users of social media. Participants are most visible when they post updates, upload content, write to each other or on each other's walls, or like/share/retweet the content others have posted. Accordingly, the data collected from APIs has an in-built bias towards those types of users that are the most 'active' content contributors, whereas the data say very little, at best, about the so-called 'lurkers' who may read their newsfeeds and Twitter streams with great interest and on a daily basis, but who barely post anything to the stream themselves (Giglietto, Rossi & Bennato 2012; Gonzáles-Bailón et al. 2013). Users only become visible in the APIs to the extent that they perform actions that are also visible for fellow users in their personal networks (i.e., connect to someone, post something). The APIs of Facebook and Twitter do not make logs of click-through patterns available for researchers to map (although such data logs are surely used by the companies themselves). At the same time, lurkers are by far the predominant group of users of social media (e.g. van Dijck 2009). This is not necessarily problematic for the use of APIs for research, but it certainly raises questions about the kinds of inferences and conclusions we can make about specific social media and their users in studies relying on data collected from APIs. For instance, if we want to

know more about lurker behavior, perhaps the only possible way of getting relevant data is to ask the lurkers themselves. Similarly, random sampling from the API is not necessarily the most useful entry point for studying 'typical users' – these must be identified based on other methods (e.g. national representative surveys of social media users, such as those of the Pew Internet and American Life Project or the Oxford Internet Survey), and sampled strategically.

In relation to the discussion of generalization, one key methodological issue concerns sampling and associated questions of representativeness. Whereas it is possible to evade the problem of ensuring a representative sample by simply sampling all users of Facebook or Twitter, as done by Kwak and colleagues (2010) in their study of Twitter, for example, this solution requires an immense server capacity, or server-side access, and is hardly an option available to most researchers.

As recently argued by Herring (2010), it is difficult (if not impossible) to estimate the quality and representativeness of a sample if you don't know the population from which the sample is drawn. For example, what is a random sample of tweets collected from the public timeline at a given point in (or period of) time on Twitter a sample of? As pointed out by Neuhaus and Webmoor (2012), many social media companies do not release more than sample data through their APIs (but for a decent sum of money, it is possible to get full access to the Twitter Firehose (i.e. the full Twitter database), for example). Exact knowledge of the population needed to construct a representative sample from which to draw statistical generalizations requires server-side access, which most of us do not have.

A few studies have sought to examine sample quality from the publicly available search and streaming APIs of Twitter. For instance, Gonzáles-Bailón and colleagues found that the often-used Twitter search API is limited in terms of systematically collecting data produced by peripheral users of the service (Gonzáles-Bailón et al. 2013: 15-16), owing to how the API filters data from the database. Morstatter and colleagues (2013) compared a sample from the Twitter streaming API with the Twitter Firehose and found that in particular smaller samples collected from the streaming API tend to misrepresent the volume of hashtags when compared to the Firehose, with implications for studies of hashtags and trending topics on Twitter. Indeed, the APIs of social media companies appear to offer restricted access to their underlying databases without providing thorough documentation of how the API filters the database (Morstatter et al. 2013).

The issue of sampling is less a problem for qualitative research, which samples purposively with a view to saturation and richness rather than representativeness. Nonetheless it is important to reflect upon the status and possible sampling biases in any piece of research using behavioral data from APIs. Research using APIs should always include a critical assessment of the sample and the in-built limitations to generalization when reporting findings. The explicit address of basic sampling biases creates transparency and thereby enhances the credibility of the empirical study.

Validity and data quality: Asking 'What' rather than 'why'

The large amount of data to be collected through the APIs may indeed inform us about the qualitative question 'what' users are doing, but does not say much

about 'why' users are doing what we can see they are doing (cf. Mahrt & Scharkow 2013). Along these lines, another challenge of the use of APIs in generating rich understandings of social media use is epistemological and has to do with the nature and quality of behavioral data. While such data are highly valid and reliable in terms of tracking various activity patterns, the behavioral data say quite little about the meanings that users ascribe to their social media use, and the contexts of use. Like any tool or method, working with APIs thus delimits the types of research questions that can be asked and answered. In terms of validity, other methods are necessary if we wish to contextualize social media usage in broader patterns of communication and relate behavior to underlying patterns of meaning-making. Hence, behavioral logs may be complemented with other types of data, for instance, interviews and observations, to get a richer, contextualized understanding of social media use.

One additional, crucial aspect of the assessment of validity of data collected from the API is the nature of basic user data entered to social media profiles. Such data is self-reported, which may distort the validity of the data set, for instance, if the user enters false information, or when fake profiles and spambots are included (cf. Karpf 2012). However, the validity of self-reported data is not unique to research based on data from APIs, but is a general challenge for empirical research.

Reliability: Software and API design structures

Related to the issue of sampling and representativeness, and the lack of transparency in API data, a third challenge in using APIs for research concerns reliability. The usefulness of APIs for researchers is very much dependent on the

developers and commercial providers of the service. In essence, the developers and service providers decide what data are interesting and relevant, as they control what is made available for analysis to developers and researchers. They can freely decide if they only want to put samples of the total user activity in the API and if they want to censor out specific types of data, for instance, status updates with words such as terrorism and Anonymous. Hence, the reliability of data gathered through the API is difficult to test for the researcher, and companies continuously make changes to the API. For instance, if the companies can commercialize on specific sets of data, they will likely remove these data from the API as seen with Twitter's increasingly restrictive API access following the explosive user growth of the service (Gonzáles-Bailón et al. 2013). Along these lines, in the case of Facebook (February 2013), the API category 'application' includes third party applications such as YouTube, but does not include Spotify and Pinterest, companies that have a special (functional and commercial) agreement with Facebook (Bechmann and Vahlstrup 2013). Ongoing changes to the APIs add further to the lack of transparency of the API data that researchers use. Researchers struggle to keep track with the development of the APIs, and previously solid and well-tested data collection strategies and software scripts for accessing the APIs may become obsolete.

Another limitation to the use of APIs to collect data for research purposes is that both Facebook and Twitter at the time of writing have a maximum number of times per hour that the researcher can call the API. Hence, there are significant risks of missing data without being able to detect it and thereby assess the reliability of the data set, when using the API to collect data. In that

sense, the API is a fragile and unstable entry point for the data collection for researchers.

As researcher we tend to assume that data drawn trough APIs are reliable, but using only this method for data collection we cannot test if reliability is actually high. It is, at best, difficult to assess whether all relevant data have been collected or if in fact there are blind spots, owing to server down time or censorship, for instance (Bechmann and Vahlstrup 2013). To judge the reliability of the data archive, in the case of the qualitative Facebook and Twitter studies carried out by Bechmann (2014, forthcoming) and Lomborg (2011), full manual checks of the sampled profiles were performed – a procedure that is hardly viable for larger-scale data sets. The difficulty of ensuring data reliability when work with API data has implications for theory development on social media use. As demonstrated by Gonzáles-Bailón et al. (2013), researchers should be careful when inferring theoretical propositions from data collected through APIs, because these are likely to be distorted.

For qualitative studies, where the reliability of API data can be manually checked to ensure that nothing is missing, a set of other reliability issues arises, concerning the archived textual corpus of social media use (Lomborg 2012). In the translation, or reduction, of actual communicative behaviour on a given site to a textual data archive or database, what happens? As Brügger (2011) argues, the web archive is not a 1-1 copy of what was actually on the live web at a given point in time. Textual or graphical archives look different from the interface and abundance of impressions that the user will meet when accessing Twitter or Facebook. This urges us to think of social media data archives not as reliable reproductions of user behaviour, but as actively constructed research objects

that are highly structured, and rely on technical constraints, research choices, as well as the tools used for data display.

Legal and ethical issues

On top of the methodological issues of generalization, validity, and reliability, there are legal and ethical implications of using data collected through APIs as part of research. A central legal and ethical aspect of API research is the use of content that may be considered private by the participants. On services such as Facebook content is technically semi-private even though it tends to be more public by default (Stutzman, Gross & Acquisti 2012). On Twitter content is by default (technically) public. Legally (in the EU at least), even technically public social media data can be personal and sensitive data and must be handled according to privacy laws. According to the EU directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data (EU 1995: Article 2a) personal data is:

> *any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.*

That is, in EU countries, API data would be considered personal and must be handled accordingly. In other parts of the world, API data may have a different

status, as personal data are viewed upon differently across national and cultural contexts (Ess 2009).

Even though postings on social media such as Twitter and Facebook may include personal and identifying information (name, date of birth, occupation, etc.), it is unclear to what extent the personal information is of a sensitive nature. Most of the data appears to be highly mundane (i.e. everyday musings, tastes and small talk). However, the researcher does not know at the time of retrieval if seemingly mundane data will come to contain highly sensitive information at a later point in time. The lack of clarity as to the sensitivity of personal data demands that researchers should be careful concerning the legal procedures of data collection and handling, but also calls for ongoing reflection and transparency concerning the ethical procedures and choices that are part of the research project.

At the same time the research community has to accept that the legal procedures are executed differently for different regions (e.g. Institutional Review Boards and National Data Protection Agencies) with a potentially different focus on the possible threats to participants in social media research projects. These differences may create obstacles for cross-cultural research on specific social media services, drawing on data from APIs. Accordingly, researchers may have to recognize and adjust to different interpretations of what is sensitive data in different regions, for example, sexual orientation, religion, and politics (Ess 2009; see also Bechmann 2014). Furthermore, the legal and ethical guidelines of different regions are typically modelled in response to quantitative approaches. This may make rich, contextualized API datasets for

qualitative analysis difficult to collect and handle according to legal and research ethical requirements.

Concerning ethics, the use of APIs as research tools is characterized by a high degree of access to the user data patterns, which may lead to an unwanted and unforeseen exposure of private data for the participants. This raises a set of key ethics questions to be considered. For instance, how can we ensure that participants are adequately informed and protected when collecting data from APIs? How much of the identifiable data can we cite or use as illustrative patterns when publishing? We contend there are no final answers to these questions. Rather, they must be carefully considered in the context of any concrete research project using APIs.

As argued by the Association of Internet Researchers (Ess and AoIR 2002; Markham and Buchanan 2012), ethical issues may vary with the research question, scope of data collection, and whether the intended analysis is centered on qualitative, in-depth examination of individual users, or is based on quantitative analysis of aggregates of users. That is to say, ethical judgment must be based on the case at hand, rather than one-size-fits-all criteria. At the same time, however, two general challenges concerning research ethics arise from using APIs, namely the questions of informed consent and anonymization of data in the processes of collecting and analyzing data and publishing findings.

Informed consent

The issue of informed consent in social media studies has two layers: one concerns the informed consent that users give to the service providers, the other concerns the specific use of personal data for research purposes.

As Neuhaus and Webmoor (2012) argue, when users sign up for a social media service, they have to accept the terms of use, often implying letting the service provider use their information to improve and monetize their product. Researchers make use of this informed consent when accessing the data through the API. On Facebook the researcher need to ask for additional permission to collect data whereas this is not the case on Twitter, owing to their statuses as semi-private and public by default services.

On top of the consent granted to the service provider, researchers would normally be required to make an informed consent agreement with the users to participate in a research project (e.g. as requested by IRBs in the U.S. or National Data Agencies in the EU). That is to say, the issue of whether to ask participants for their consent to a research project does not per se disappear just because the data is made publicly available by the service provider and can be collected by a 'hidden observer'. To be sure, whether to ask for participants' permission before retrieving their data from the APIs of social media depends on the research purposes and types of analysis involved, but a few general points can be made.

Even though API research often involve studying mundane everyday communication, the construction and analysis of social media logs over an extended period constitutes a challenge, especially for qualitative research which may involve a close tracking of identifiable persons' online activities: who they talk to, to whom they are connected, their cultural tastes and preferences, and so on. We may find cases where researchers prove to know more about the activity of the users and the public visibility thereof than the users themselves do. Indeed, the *accumulated data* on user behaviour over time may be experienced as highly personal, because it enables a close-up analysis of individual users.

Hence, in order to respect human subjects' perceived privacy, we contend that that seeking informed consent before collecting data would – at least in qualitative studies – is not only a legal matter, but also advisable from an ethical point of view (cf. Bechmann forthcoming; Lomborg 2012).

However, asking for informed consent is not a straightforward matter, even for qualitative research on communicative patterns and practices. Even when sampling from a small number of profiles, many secondary persons (e.g. friends of the Facebook participants, and conversation partners of Twitter profile holders) are involved and figure in the data archives. Asking all of these users for permission would simply be impracticable. But if secondary participants are included without explicitly asking for their consent, it becomes the researcher's responsibility to assess possible privacy problems on behalf of these users and only use the data as documentation to the extent that other measures of privacy protection, such as anonymization, are put to use, and respect of human subjects ensured.

Asking for informed consent is a viable ethical procedure for small-sample, qualitative studies using APIs, where researchers may keep close contact with the research participants, but is hardly an option for large-scale research with thousands of users involved. In quantitative studies using large data sets there is not per see any direct contact between researcher and research participant, thus further complicating the ethical conduct of research. At the same time, as we have seen in the literature review, most quantitative studies using APIs are interested in structural analysis, pattern recognition and prediction and not in single user profiles, in contrast to qualitative studies. This creates research scenarios where it may be appropriate not to seek informed consent, simply

because there is a greater distance between the analysis being made and the actual users involved in the data sets (i.e., the individual user is tracked less closely and with fewer data points, whereas as the number of users involved is increased). Instead, the legal and ethical challenges in quantitative studies using APIs often revolve around how data is anonymized both to the researcher and when presenting results.

Anonymization

Whereas the question of informed consent is principally raised in the process of collecting data, anonymization specifically ties in with data processing and publication. In qualitative studies, anonymization is difficult as the purpose of these studies is to dig deeper into the material and understand one single profile in relation to another. This would often involve looking at pictures, reading status updates and friends' comments and sometimes contextualizing this activity in relation to the broader everyday life of the user. Anonymization of the participant in the study, when presented to the public, has always been the standard in qualitative studies, but because of the contextual approach it may be impossible if even desirable to anonymize users. In addition, on publicly accessible social media such as Twitter, simple string search of text bits from tweets used in analysis might expose the author's identity (cf. Lomborg 2013).

In quantitative studies relying on large databases, researchers anonymize their datasets per default and test units are not publicly searchable, but as for instance discussed by Zimmer (2010) and Ohm (2010), it may be possible to de-anonymize individuals, not only through publicized text strings, but also social graphs from the social media service or descriptions of the dataset.

The challenges associated with anonymizing data collected from the API have implications not only for the sharing of datasets among researchers, but also for the kinds of analysis published based on this material. For example, in network analyses of structural patterns of connectivity and conversation amongst users, the information and analytic depth that may be contained in disclosing the users behind individual nodes must be weighed against the risk of violating their privacy expectations.

Conclusion: Tracking meaningful data paths from social media APIs

We have raised a critical discussion of some of the key methodological opportunities and challenges, as well as ethical considerations regarding the use of APIs in social media research. Whereas research ethical discussions of this type of research is currently receiving a bulk of attention (e.g. boyd and Crawford 2012; Markham and Buchanan 2012; Neuhaus and Webmoor 2012), a methodological discussion of the crucial dimensions of validity, reliability and generalization involved in this type of social media research is still sparse.

With this article, we have addressed sampling challenges and the associated implications for generalization from API data, questions of validity and data quality considering the behavioral and de-contextualized nature of data drawn from the APIs of social media, and challenges regarding the reliability and trustworthiness of the API in collecting all relevant data points and creating an archive or database for cleaning, processing, and analyzing the API data. Each of the points raised deserve further attention. This article intends to serve the function of opening the floor for discussion of how to proceed in a

methodologically and ethically sound way in internet studies using APIs as a research tool. Future discussion should include other APIs than the ones of Twitter and Facebook, which have been our main examples here. Along these lines, a mapping and comparison of various types of APIs may be helpful in specifying promises and pitfalls of using different APIs for research (e.g., Gonzáles-Bailón et al. 2013; Morstatter et al. 2013).

At present, APIs are primarily used for large-scale quantitative analysis of behavioral data, probably owing to the abundance of new data available for automated collection and analysis at an unprecedented scale. This article has shown a glimpse of the potential and the challenges that APIs also hold for qualitative research. Further methodological inquiry may consider the relationship between quantitative and qualitative analysis of digital trace data collected from APIs. In order to get hold of the large amounts of data quickly generated in studies using APIs, compared to other qualitative methods such as interviews, researchers may more often be urged to deploy quantitative techniques for analysis and quantify research questions, for example, 'How many of the participants use Facebook from a smartphone?' or 'At what time of the day does the participant most often status update or tweet?' Such descriptive, quantitative analyses are often necessary to gain an overview of the datasets. At the same time, possibilities for quantification should not overshadow classic qualitative analysis just because larger datasets are easy to collect. Instead, descriptive quantitative analysis could form part of the initial analytic groundwork, assisting qualitative researchers in 'opening' up the collected data and identifying relevant and meaningful questions for qualitative inquiry. The use of APIs and other computational techniques may thus prompt a further

integration of quantitative and qualitative analysis in mixed-methods research designs, to better understand social media as contemporary communicative phenomena.

References

Ausserhofer, J. & Maireder, A. 2013. National politics on Twitter. *Information, Communication & Society*, DOI:10.1080/1369118X.2012.756050

Backstrom, L., P. Boldi, M. Rosa, J. Ugander and S. Vigna 2011. *Four Degrees of Separation*. http://arxiv.org/abs/1111.4570 (Accessed April 10, 2013).

Bakshy, E., Rosenn, I., Marlow, C. Rosenn and Adamic, L. 2012. The Role of Social Networks in Information Diffusion. *Proceedings of the 21st international conference on World Wide Web*: 519-528.

Baron, N. 2008. *Always On: Language in an Online and Mobile World*. New York: Oxford University Press.

Bechmann, A. and S. Lomborg. 2012. Mapping actor roles in social media: Different Perspectives on Value Creation in Theories of User Participation. *New Media & Society*. Online first, DOI: 10.1177/1461444812462853.

Bechmann, A. and P. Vahlstrup. Forthcoming, 2013. Digital Footprints: A system for studying private user data on Facebook, Proceedings of CHI'13, 27th April – 2nd May Paris, France. ACM.

Bechmann, A. 2014, forthcoming. Managing the interoperable self, *The Ubiquitous Internet* (ed. Bechmann A & Lomborg, S), NY: Routledge.

Bernstein, M., E. Bakshy, M. Burke and B. Karrer 2013. Quantifying the invisible audience in social networks, *CHI: ACM Conference on Human Factors in Computing Systems*, April 27–May 2, Paris, France.

Berry, D. M. 2011. The computational turn. Thinking about the digital humanities. *Culture Machine*, 12. http://www.culturemachine.net/index.php/cm/article/view/440/470

Black, A., C. Mascaro, M. Gallagher, M and S. P. Goggins 2012. Twitter zombie: architecture for capturing, socially transforming and analyzing the twittersphere. *GROUP'12. Proceedings of the 17th ACM international conference on Supporting group work*, 229-238.

boyd, d. and K. Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5): 662-679.

Brügger, N. 2011. Web archiving – between past, present, and future. In *The Blackwell Handbook of Internet Studies*, ed. M. Consalvo and C. M. Ess, 24-42. Oxford, UK: Wiley-Blackwell.

Bruns, A. and J. Burgess. 2012. Researching News Discussion on Twitter: New Methodologies. *Journalism Studies*, online first: DOI: 10.1080/1461670X.2012.664428.

Bruns, A. and Y. E. Liang. 2012. Tools and methods for capturing Twitter data during natural disasters. *First Monday* 17(4).

Bollier, D. 2010. *The Promise and Peril of Big Data*. http://www.thinkbiganalytics.com/uploads/Aspen-Big_Data.pdf (Accessed March 1, 2012).

Courtois, C., P. Mechant and L. Marez. 2011. Teenage uploaders on YouTube: Networked public expectancies, online feedback preference, and received on-platform feedback. C*yberpsychology, Behavior, and Social Networking* 14(5): 315-322.

De Smedt, T., and W. Daelemans 2012. Pattern for Python. *Journal of Machine Learning Research* 13(1), 2063-2067.

van Dijck, J. 2009. Users like you? Theorizing agency in user-generated content. *Media, Culture & Society* 31(1): 41-58.

Ess, C. M. 2009. *Digital media ethics*. Cambridge, UK: Polity Press.

Ess, C. M. and AoIR Ethics Working Committee. 2002. *Ethical Decision-making and Internet Research: Recommendations from the AoIR Ethics Working Committee*. http://www.aoir.org/reports/ethics.pdf (Accessed June 9, 2012).

EU [the European Parliament and the Council]. 1995. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, *Official Journal L 281*, 23/11/1995: 31-50.

Fuchs, C., K. Boersma, A. Albrechtslund and M. Sandoval (Eds.). 2012. *Internet and Surveillance: The Challenges of Web 2.0 and Social Media*. New York: Routledge.

Gawk 2011. http://www.gnu.org/software/gawk/ (Accessed March 3 2014).

Gephi 2011. http://gephi.org (Accessed March 3 2014).

Giglietto, F., L. Rossi and D. Bennato 2012. The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and Youtube as a Research Data Source, *Journal of Technology in Human Services*, 30(3-4): 145-159.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2012). Assessing the Bias in Communication Networks Sampled from Twitter. *SSRN: http://ssrn.com/abstract=2185134*.

Hansen, D.L., B. Shneiderman and M.A. Smith 2011. *Analyzing social media with NodeXL. Insights from a connected world*. Burlington, MA: Elsevier.

Herring, S. C. 2010. Web Content Analysis: Expanding the Paradigm. In *The International Handbook of Internet Research*, ed. J. Hunsinger, M. Allen and L. Klastrup, 233-250. Berlin: Springer Verlag.

Honeycutt, C. and Herring, S. C. 2009. Beyond microblogging: Conversation and collaboration via Twitter. *Proceedings of the Forty-Second Hawaii International Conference on System Sciences (HICSS-42).* Los Alamitos, CA: IEEE Press.

Horan, T. J. 2012. 'Soft' versus 'hard' news on microblogging networks. *Information, Communication & Society*, online first: DOI:10.1080/1369118X.2011.649774.

Huberman, B. A., Romero, D. & Wu, F. 2009. Social networks that matter: Twitter under the microscope. *First Monday vol. 14(1).*

Java, A., Song, X. Finin, T. and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. *Proceedings of WebKDD/SNA-KDD '07*. ACM Press.

Karpf, D. 2012. Social science research methods in internet time. *Information, Communication & Society*, 15(5): 639-661.

Kramer, A.D.I. 2012. The spread of emotion via facebook. *Proceedings of CHI2012*, ACM: 767-770.

Kwak, H., C. Lee, H. Park and S. Moon. 2010. What is Twitter, a Social Network or a News Media? *Proceedings of the International World Wide Web conference*

(WWW 2010): 591-600.

Leskovec, J. and E. Horwitz. 2007. *Planetary-Scale Views on an Instant-Messaging Network. Microsoft Research Technical Report.* Microsoft.

Lomborg, S. 2011. *Social media. A genre perspective*. PhD Thesis. Aarhus University.

Lomborg, S. 2012. Researching communicative practice: Web archiving in qualitative social media research. *Journal of Technology in Human Services, 30(3/4)*: 219-231.

Lomborg, S. 2013. Personal internet archives and ethics. *Research Ethics*, 9(1): 20-31.

Lotan, G., E. Graeff, M. Ananny, D. Gaffney, I. Pearce and d. boyd. 2011. The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communications* 5: 1375-1405.

Mahrt, M & M. Scharkow 2013. The Value of Big Data in Digital Media Research, *Journal of Broadcasting & Electronic Media*, 57(1): 20-33.

Markham, A. 2012. Moving into the flow: Using a network perspective to explore complexity in Internet contexts. In *Network analysis - methodological challenges*, ed. S. Lomborg, 45-58. CFI Monograph series 14, Aarhus University.

Markham, A. N. and E. A. Buchanan. 2012. *Ethical Decision-Making and Internet Research (version 2.0) Recommendations from the AoIR Ethics Working Committee.*

McChesney, R. 2007. *Communication Revolution: Critical Junctures and the Future of Media*. New York: The New Press.

Morstatter, F., Pfeffer, J., Liu, H. and Carley, K. M. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *ICWSM 2013*.

Naaman, M., Boase, J. and Lai, C.-H. 2010. Is it really about me? Message content in social awareness streams. *Proceedings of CSCW-2010*: 189-192.

Neuhaus, F. and T. Webmoor. 2012 Agile ethics for massified research and visualization. *Information, Communication & Society* 15(1): 43-65.

Ohm, P. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, *UCLA Law Review*, 57, pp. 1701-1777.

OpenRefine 2012. http://openrefine.org (Accessed March 3 2014).

Reips, U.-D. and P. Garaizar 2011. Mining Twitter: Microblogging as a source for psychological wisdom of the crowds. *Behavior Research Methods* 43, 635-642.

Rogers, R. 2012. *Digital Methods*. Cambridge, MA: MIT Press.

Serrano, A. 2011. The social media explosion: By the numbers, *Fiscal Times*, September 12. http://www.thefiscaltimes.com/Articles/2011/09/12/The-Social-Media-Explosion-By-the-Numbers.aspx#page1. (Accessed July 6, 2012).

Smith, M. A., B. Schneiderman, N. Milic-Frayling, E. M. Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer and E. Gleave 2009. Analyzing (social media) networks with NodeXL. *C&T '09. Proceedings of the fourth international conference on Communities and Technologies*, 255-264.

Stutzman, F., R. Gross and A. Acquisti 2012. Silent Listeners: The Evolution of Privacy and Disclosure on Facebook*, Journal of Privacy and Confidentiality*, 4(2), p. 7-41.

Vicente, C. R., I. Assent and C. S. Jensen. 2011. Effective Privacy-Preserving Online Route Planning. *12th IEEE International Conference on Mobile Data Management (MDM)*: 119-128. IEEE Computer Society Press.

Weller, K. and C. Puschmann 2011. Twitter for Scientific Communication: How Can Citations/References be Identified and Measured? *Proceedings of the ACM WebSci'11*: 1-4, June 14-17 2011, Koblenz, Germany.

Wu, S., W. A. Mason, J. M. Hofman and D. J. Watts. 2011. Who Says What to Whom on Twitter. *Proceedings of the International World Wide Web Conference (WWW 2011):* 705-714.

Yahoo! Pipes 2007. http://pipes.yahoo.com/pipes/

(Accessed March 3 2014).

YourTwapperKeeper 2011. https://github.com/540co/yourTwapperKeeper (Accessed March 3 2014).

Zimmer, M. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics & Information Technology* 12(4): 313-325.

Notes

[1] Even though Facebook has many public profiles that can be crawled a majority of Facebook users have private profiles with customized privacy settings that only makes content visible to a selected group of friends or limited network. In contrast, Twitter is public by default, that is, most Twitter profiles are publicly accessible online, and do not require readers to identify themselves through a login and password to access the posted tweets. As cases in this article, Facebook and Twitter thus represent a semi-private and a public social media service, respectively. Furthermore, communication on Twitter is in a sense more distributed than on Facebook, as any given user's tweet is archived on the user's own profile, whereas on Facebook users can write on one another's profiles.

[2] code.google.com/p/tatiana/

[3] www.digitalfootprints.dk

[4] Digital Footprints uses Json Array and the data collected from Facebook was imported in the Digital Footprints local server database as multiple data points. These data was sorted according to interface categories and supplemented with different statistical functions, fuzzy search and graphical views.